Original article

A hybrid method for performance analysis of *G/G/m* queueing networks

Boualem Rabta

Entreprise Institute, University of Neuchâtel, Rue A.L. Breguet 1, CH-2000 Neuchâtel, Switzerland

Abstract

Open queueing networks are useful for the performance analysis of numerous real systems. Since exact results exist only for a limited class of networks, decomposition methods have been extensively used for approximate analysis of general networks. This procedure is based on several approximation steps. Successive approximations made in this approach can lead to a considerable error in the output. In particular, there are no general accurate formulas for computing the mean waiting time and the inter-departure variance in general multiple-server queues. This causes the results from decomposition methods when applied to *G/G/m* queueing networks to be very approximative and to significantly deviate from actual performance values. We suggest substituting some approximate formulae by low-cost simulation estimates in order to obtain more accurate results when benefiting from the speed of an analytical method. Numerical experiments are presented to show that the proposed approach provides improved performance.
© 2013 IMACS. Published by Elsevier B.V. All rights reserved.

1. Introduction

Queueing networks are an extremely useful class of models that have seen use in a host of application areas. In particular, they have been successfully used to model the performance of a variety of complex systems such as computer systems, communication networks, production lines and manufacturing systems. Queueing models allow the consideration of the randomness of different components. Unfortunately, exact results exist only for a limited class of networks (product form). Decomposition methods among other approximation methods, have been extensively used to obtain approximate results. This approach has been developed by Kuehn [23], Whitt [34] among others, and is based on two-moments approximations, that is, all stochastic processes are characterized by their mean and squared coefficient of variation (SCV) (see, e.g., [26]). Whitt [34] proposed the extension of this method to multiple server nodes by suggesting approximate formulas for computing the mean waiting time and the inter-departure second moment. However, a drawback of this method is that it performs well in some situations, but not others (see, e.g. [35]). Whitt [37] proposed an enhancement to the parametric-decomposition method. Instead of using a variability parameter for

E-mail addresses: boualem.rabta@unine.ch, brabta@yahoo.fr

each arrival process, he suggested using a variability function; i.e., the variability parameter should be regarded as a function of the traffic intensity of a queue to which the arrival process might flow.

The classical decomposition method assumes that all arrival processes to a station are renewal processes, ignoring therefore the possible correlation among different arrival streams. Also, it is assumed that the superposition of renewal processes is a renewal process which is generally not correct. Albin [2] remarks that if at least one component process is not Poisson, then the superposition process is not renewal and the intervals between arrival points are identically distributed, but are not independent. Ignoring this interdependence (i.e. correlation) between inter-arrival intervals might cause errors in the approximation. Kim et al. [18] proposed the innovations method as an improvement to Whitt's method by replacing relations among squared coefficients of variability with approximate regression relationships among in the underlying point processes. These relationships allow adding information on correlations between different streams. Kim [16] combined the innovations method with Whitt's variability functions to deal with the heavy traffic bottleneck phenomenon.

Van Nyen et al. [33] pointed out that the classical decomposition method for networks with multiple customer classes and aggregation as proposed by Albin [1,2] and Whitt [34] performs poorly in manufacturing context. They used simulation to show that the method generates serious approximation errors in some cases. Note that since the Albin–Whitt procedure has been proposed, several improvements followed (e.g., multi-class decomposition methods proposed by different authors (e.g., Bitran and Tirupati [6], Whitt [36], Caldentey [9])) and many papers proposed methods to capture correlations in the arrival process reported as the principal source of errors by Van Nyen et al. (see, e.g., Heindl and Telek [14], Heindl et al. [15], Kim [17,16], Kim et al. [18], Balcioglu et al. [7]).

Ignoring the correlation among arrival streams is not the only source of errors in the decomposition method. This procedure includes a number of approximation formulae that are more or less precise depending on the input values, and generates errors at each step. Previous papers focused on improving the merging step (e.g., [18,7]) and improving the estimation of inter-departure SCV (e.g., in multiple customer classes case [6,36,9]). Haverkort [13] proposes to approximate $G/G/1$ nodes by $PH/PH/1$ ones through approximating arrival and service distributions by phase-type distributions based on their first two moments and then deriving exact measures for each node using matrix-geometric techniques. This suggestion leads to improved performance but is only limited to single-server nodes. In addition, the quality of the approximation may also depend on higher moments. The inter-departure SCV computation by means of Marshall's formula in the classical decomposition method, involves the computation of the waiting time. For the $G/G/1$ queueing system only approximate formulae for the waiting time are available and even less are available for the multiple-server system $G/G/m$. These formulae seem to perform more or less accurately depending on the system's parameters.

Simulation is perhaps the most popular approach to the performance measure of complex systems. For instance, Cruz et al. [11] discuss the use of simulation for performance evaluation of mobile communication networks and Cruz et al. [12] describe a simulation model for state-dependent finite queueing network analysis. It is identified that simulation models allow a higher level of realism and system's details but they can be cumbersome to optimize (much time to build the model and run different scenarios), and their accuracy is largely dependent on the quality of the calibration data. In particular, simulation of large systems can be expensive both in terms of CPU time and use of available resources (e.g., memory, processors).

In this paper, we propose a hybrid simulation-decomposition method for the analysis of $G/G/m$ queueing networks. We aim, to show that the improvement of the classical decomposition method is possible and to propose a more precise tool. A low cost simulation algorithm based on a set of recursive equations proposed by Krivulin [22], is used to compute the inter-departure SCV instead of the approximative formula used in the classical decomposition method. Furthermore, the same set of recursive equations allows us to simulate the waiting time for $G/G/m$ nodes. This allows us to obtain more accurate results when considering real service distributions and multiple-server nodes.

The proposed approach attempts to combine the speed of the analytical decomposition procedure with the precision of simulation. Shanthikumar and Sergent [30] pointed out that the estimators obtained from the hybrid simulation/analytic models have lower variance than the variance of the estimators of the traditional simulation models. As opposed to a pure simulation method, our method is faster and less expensive in terms of computer resources since no memory allocation is required for entities (queues, servers, etc.). Numerical experiments demonstrate that improvements are made in specific situations and that performance measures such as the waiting time at the bottleneck stations and overall cycle time obtained by means of the hybrid method, are more accurate.